3B3-NFC-05a-1in1

企業名に関する関心動向のトピックモデリングを用いた日中市場シェアの分析

Analyzing Japanese and Chinese Market Share based on Topic Modeling of Concerns on Company Names

宇津呂 武仁[*1] Takehito Utsuro　徐 凌寒[*2] Linghan Xu　聶 添[*2*3] Tian Nie　趙 辰[*2] Chen Zhao　李 佳奇[*2] Jiaqi Li　河田容英[*4] Yasuhide Kawada

[*1]筑波大学システム情報系 Fclty. Eng, Inf. & Sys, Univ. of Tsukuba
[*2]筑波大学大学院システム情報工学研究科 Grad. Sch. Sys. & Inf. Eng, Univ. of Tsukuba
[*3]パイオニア 商品統括部 Product Management Division, Pioneer Corporation
[*4](株) ログワークス Logworks Co., Ltd.

This paper proposes to utilize a search engine as a social sensor which is to be used for predicting market shares. More specifically, this thesis studies a task of comparing rates of concerns of those who search for Web pages among several companies which supply products, given a specific products domain. In this paper, we measure concerns of those who search for Web pages through search engine suggests. Then, we analyze whether rates of concerns of those who search for Web pages have certain correlation with page view statistics of EC sites. We apply the proposed technique to both Japanese and Chinese. The results of the analysis show that those statistics have certain correlations between each other.

表 1: クエリ・フォーカスごとのサジェスト数および収集されたウェブページ数 (日本語，2015 年 8 月 6 日収集)

クエリ・フォーカス	サジェスト数	ウェブページ数
ASUS	840	5,012
Lenovo	839	5,163
NEC	909	6,329
SONY	812	5,695
シャープ	900	5,885
パナソニック	938	6,541
三菱電機	847	5,301
富士通	896	6,071
日立	912	6,568
東芝	896	6,367
混合文書集合	—	57,582

1. はじめに

本論文では，検索エンジン・サジェストによって測定される関心事項の情報を最大限に有効活用するタスクとして，特定商品ジャンルにおける製品・サービス等の供給者である複数の企業の間で，検索における関心の度合いを比較するというタスクを設定する (図 1(日本語)，および，図 2(中国語))．そして，検索における関心の度合いが，通販サイトにおけるページビュー統計との間でどの程度の相関を持つのかについて分析を行う [今田 16]．特に，本論文では，日中二言語を対象として本手法を適用し，日中両言語において，一定以上の相関を示すという結果を報告する．

2. クエリ・フォーカス

本論文では，詳細な情報を検索したい対象を「**クエリ・フォーカス**」と呼ぶ．日本語を対象としては，クエリ・フォーカスとして，表 1 に示す電気メーカー 10 社を指定し，各種電気製品ジャンルにおける関心の割合を比較する．日本語を対象としては，これらの検索対象を $q_j (j = 1, \ldots, 10)$ とする．中国語を対象としては，クエリ・フォーカスとして，表 2 に示す電気メーカー 25 社を指定し，各種電気製品ジャンルにおける関心の割合を比較する．中国語を対象としては，これらの検索対象を $q_j (j = 1, \ldots, 25)$ とする．

表 2: クエリ・フォーカスごとのサジェスト数および収集されたウェブページ数 (中国語，2015 年 11 月 20~28 日収集)

クエリ・フォーカス	サジェスト数	ウェブページ数
三星(Samsung)	2,585	3,558
乐视(Letv)	1,963	2,312
海信(Hisense)	1,462	1,723
创维(Skyworth)	436	523
TCL	1,763	1,961
小米(Xiaomi)	2,506	3,413
飞利浦(Philips)	1,847	1,952
康佳(Konka)	1,031	1,268
长虹(Changhong)	1,302	1,514
夏普(SHARP)	1,102	1,345
索尼(SONY)	2,240	2,870
海尔(Haier)	2,106	2,672
LG	933	1,176
东芝(TOSHIBA)	1,222	1,468
ThinkPad	1,591	1,764
外星人(alienware)	1,883	2,400
Acer	907	1,153
神舟(Hasee)	1,042	1,272
DELL	1,710	2,088
msi	2,368	3,225
HP	2,188	2,991
苹果(APPLE)	2,721	3,496
华为(Huawei)	2,491	3,339
华硕(ASUS)	1,981	2,461
联想(Lenovo)	2,318	2,838
混合文書集合	-	54,591

3. サジェストを用いたウェブページの収集

日本語を対象としては，選定した評価用クエリ・フォーカスに対して，Google[*1] 検索エンジンを用いて，一クエリ・フォーカス当り約 100 通りの文字列を指定し，最大約 1,000 語のサジェストを収集する．一方，中国語を対象としては，選定した評価用クエリ・フォーカスに対して，Baidu[*2] 検索エンジンを

連絡先: 宇津呂 武仁，筑波大学システム系，
〒 305-8573　茨城県つくば市天王台 1-1-1, 029-853-6537

∗1 http://www.google.com/
∗2 http://www.baidu.com/

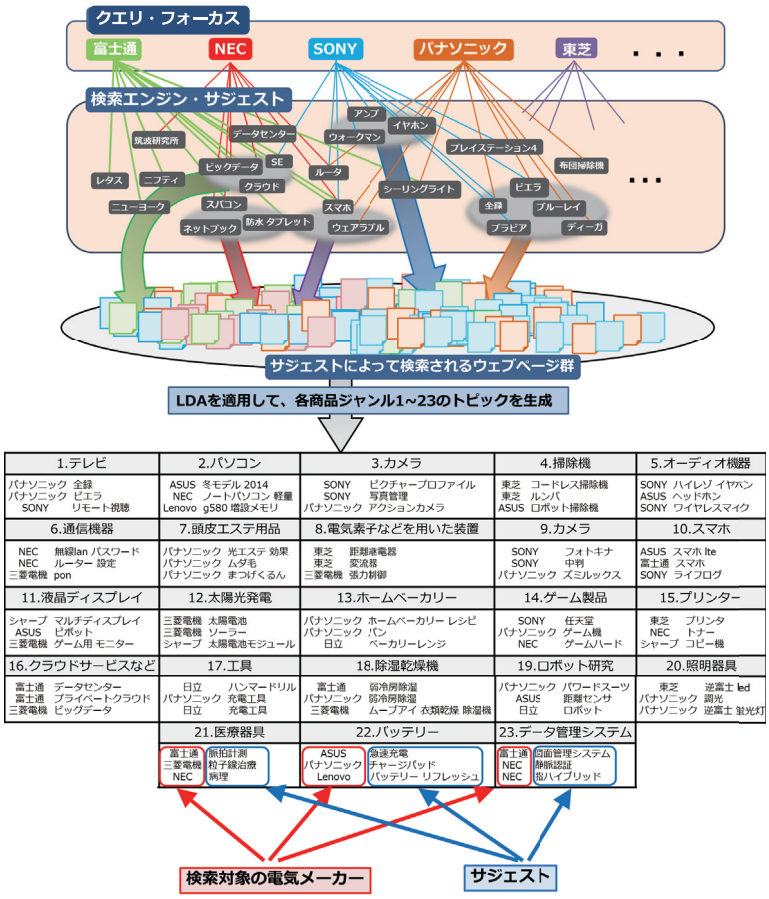

図 1: 検索における関心の割合を企業間で比較する処理の流れ (1) (日本語)

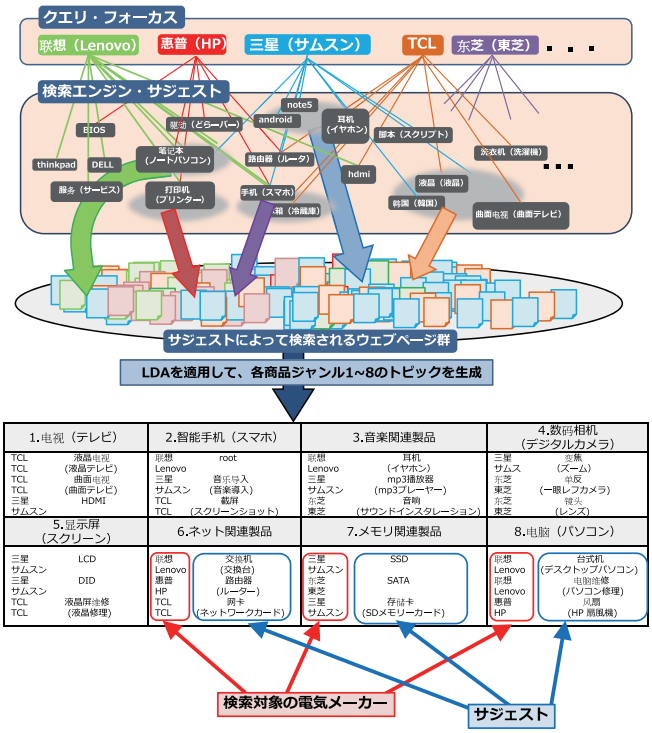

図 2: 検索における関心の割合を企業間で比較する処理の流れ (2) (中国語)

用いて，一クエリ・フォーカス当り 412 通りのピン音 (pinyin) を指定し，最大 4,120 語のサジェストを収集する．さらに，あるクエリ・フォーカスに対して収集されたサジェストの集合を S として，$s \in S$ となるサジェスト s に対して，クエリ・フォーカス q_j との AND 検索により上位 N 件以内に検索されるウェブページ d の集合 $D(q_j, s, N)$ (本論文では，日本語を対象としては $N = 10$，中国語を対象としては $N = 2$ とする) を作成する．ここで，ウェブページの収集には Yahoo! Search BOSS API[*3] を用いる．また，各企業 q_j ごとに収集したウェブページ集合 $D(q_j)$ を混合し，混合文書集合 D を作成する．各クエリ・フォーカスごとのサジェスト数およびウェブページ数の一例を表 1(日本語) および表 2(中国語) に示す．

各ウェブページは，クエリ・フォーカスおよび各サジェストの AND 検索によって検索されたものである．したがって，あるウェブページには，一つ以上のサジェストが対応することになる．各ウェブページ d に対して，$d \in D(q_j, s, N)$ となるサジェスト s を集めた集合を $S(q_j, d)$ とする．

4. トピックモデルを用いた話題の集約

本論文では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [Blei 03] を用いる．LDA を用いたトピックモデルの推定においては，語 w の集合を V として，語 $w (w \in V)$ の列によって表現された文書の集合と，トピック数 K を入力として，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)$ $(w \in V)$，及び，各文書 d におけるトピック z_n の確率分布 $P(z_n|d)$ $(n = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++[*4]を

*3 http://developer.yahoo.com/search/boss
*4 http://gibbslda.sourceforge.net/

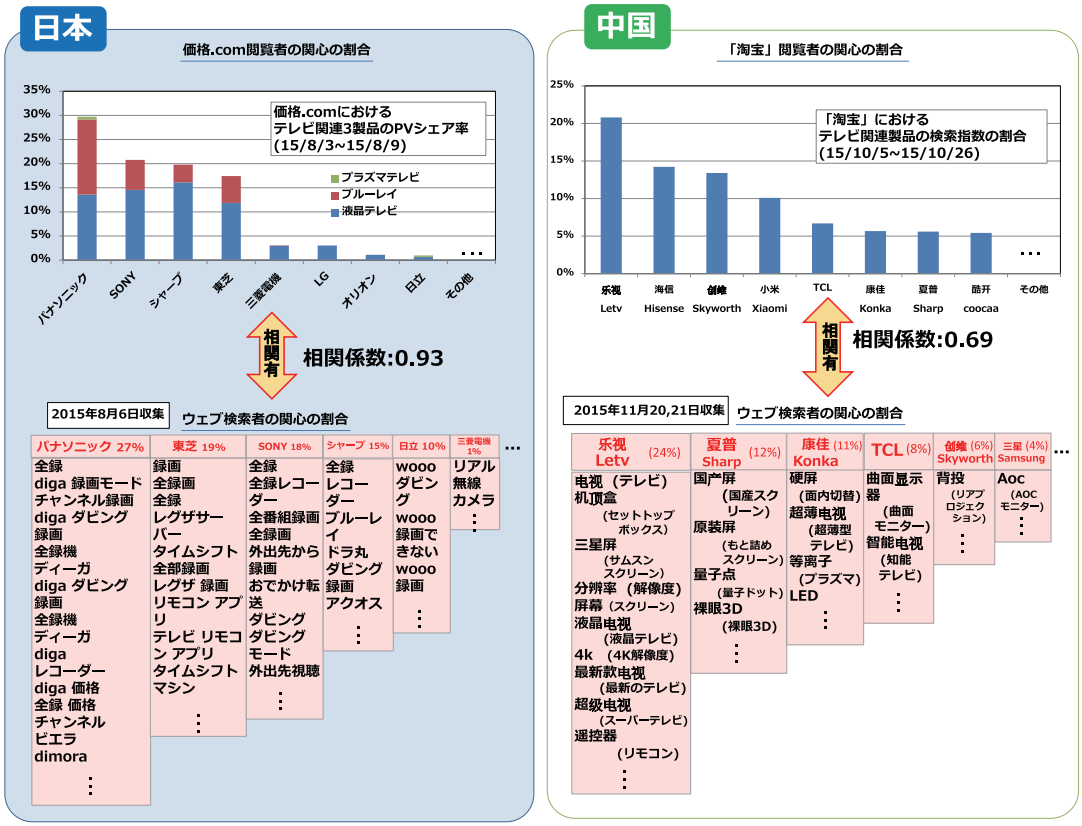

図 3: 日中における「テレビ関連製品」分野でのウェブ検索者の関心の割合と通販サイトにおける関心の割合の間の相関分析

用いた.

また，GibbsLDA++では，各トピック z_n において確率 $P(w|z_n)$ の高い順に語 w を W 件出力することができる．本論文においては，$W = 20$ として，トピックの話題分析の際に参考情報として用いている．

本論文では，各文書に対してトピックを一意に割り当てることで，各文書を分類することとした．記事集合を D，トピック数を K，1 つの文書を d $(d \in D)$ とする．文書 d におけるトピックの分布において，確率が最大のトピックに，文書 d を割り当てている．

また，各ウェブページには，トピックが対応付けられている．一つのトピックに対して割り当てられた一つ以上のウェブページに対応するサジェストを収集することにより，一つのトピックに一つ以上のサジェストが割り当てられていることになる．クエリ・フォーカス q_j に対してあるトピック z_n に割り当てられたウェブページ集合を $D(z_n, q_j)$ とすると，トピックに割り当てられたサジェスト集合 $S(z_n, q_j)$ は次式となる．

$$S(z_n, q_j) \quad = \quad \bigcup_{d \in \ D(z_n, q_j)} S(q_j, d)$$

トピック z_n の話題分析を行う際には，各クエリ・フォーカス q_j(日本語においては $j = 1, \ldots, 10$，中国語においては $j = 1, \ldots, 25$) に対する集合 $S(z_n, q_j)$ 中のサジェストのうち，全クエリ・フォーカスに対する総頻度の上位 20 個を参照することによって話題を分析する．

さらに，3. 節において作成された混合文書集合に対して，確率値 $P(z_n|d)$ の下限値を設定し，企業別のウェブページ集合および検索エンジン・サジェスト集合を作成する [今田 16]．確率値 $P(z_n|d)$ の値が下限値 θ_{lbd} 以上のウェブページを収集し，集合 $D(z_k, q_j, \theta_{lbd})$ を作成する．また，それらのウェブページに割り当てられているサジェストを収集した集合を $S(z_n, q_j, \theta_{lbd})$ とする．

5. サジェストの統計分布と通販サイトページビュー統計の間の相関の分析

本論文では，トピック数 K を 50 から 100 程度まで変化させてトピック推定を行った．これらの結果のうち，商品ジャンルとしてまとまりのよいトピックがなるべく多く観測できた結果をとりあげたものを，図 1(日本語)，および，図 2(中国語) に，それぞれ示す．図 1 の日本語での結果においては，商品ジャンルとしてのまとまりがよいトピックが 23 個観測できた．一方，図 2 の中国語での結果においては，商品ジャンルとしてのまとまりがよいトピックが 8 個観測できた．

次に，本節では，4. 節で抽出したサジェストの集合に対して，サジェスト数の企業別割合を算出する．集合 $S(z_n, q_j, \theta_{lbd})$ における検索エンジン・サジェスト数の企業別割合を次式で表す．

$$rate(z_n, q_j, \theta_{lbd}) \quad = \quad \frac{\left|S(z_n, q_j, \theta_{lbd})\right|}{\sum_i \left|S(z_n, q_i, \theta_{lbd})\right|}$$

そして，各トピック数 K において，確率値 $P(z_n|d)$ の下限値 θ_{lbd} の値を 0~0.9 の範囲で変化させて，トピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} の組について，検索エンジン・サジェスト数の企業別割合の分析を行った．具体的に，本論文では，日中両言語において,「テレビ関連製品」分野，および,「パソコン関連製品」分野に相当する各トピックにおいて，検索エンジン・サジェスト数の企業別割合の分析を行った．

まず，日本語を対象としては，検索エンジン・サジェスト数の企業別割合と，価格.com 閲覧者の関心の割合 (ページビュー統計)，価格.com における市場シェアとの間で相関係数の平均値が最大となるトピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} を求めた結果,「テレビ関連製品」分野においては $K = 90$, $\theta_{lbd} = 0.3$,「パソコン関連製品」分野においては $K = 90$,

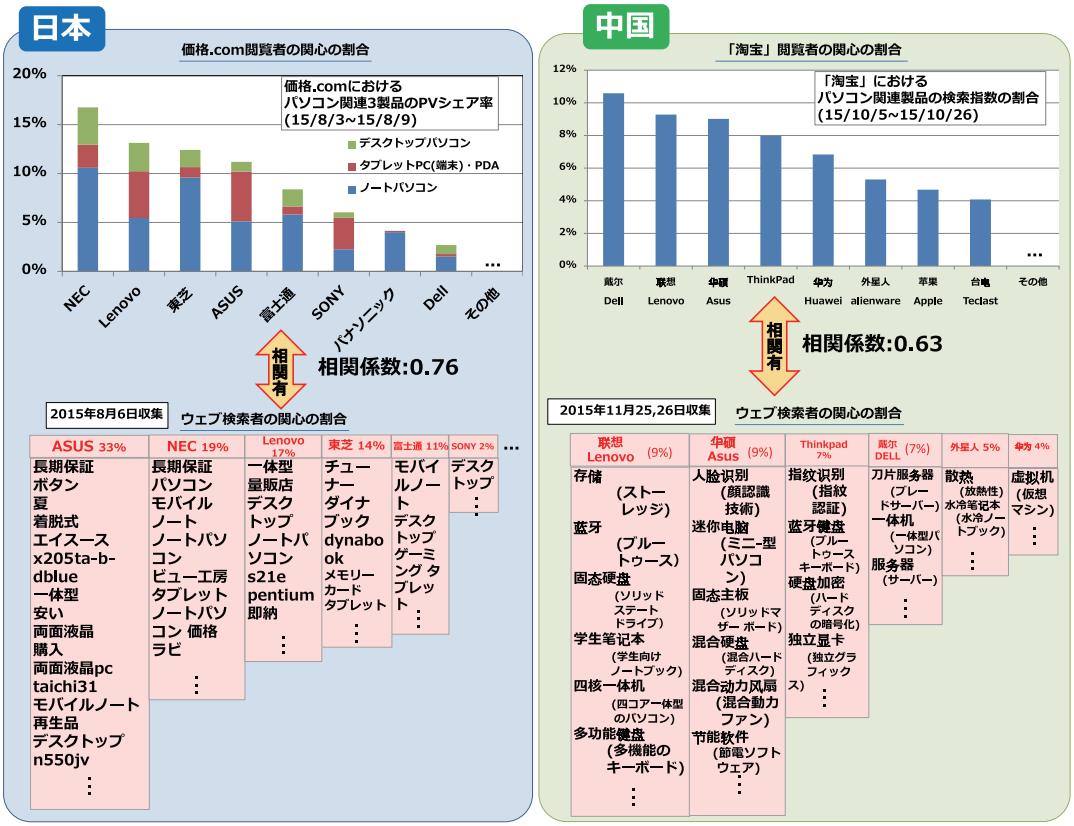

図 4: 日中における「パソコン関連製品」分野でのウェブ検索者の関心の割合と通販サイトにおける関心の割合の間の相関分析

$\theta_{lbd} = 0.4$ となった．一方，中国語を対象としては，検索エンジン・サジェスト数の企業別割合と，中国最大の通販サイト「淘宝」[*5] における各製品分野ごとのページビュー統計[*6] との間で相関係数が最大となるトピック数 K および確率値 $P(z_n|d)$ の下限値 θ_{lbd} を求めた結果，「テレビ関連製品」分野においては $K = 50$，$\theta_{lbd} = 0$，「パソコン関連製品」分野においては $K = 70$，$\theta_{lbd} = 0$ となった．これらのうち，「テレビ関連製品」分野および「パソコン関連製品」分野の各々において，検索エンジン・サジェスト数の企業別割合，および，通販サイトにおけるページビュー統計を比較し，相関係数を算出した結果を，それぞれ，図 3，および，図 4 に示す．この結果のうち，特に，「テレビ関連製品」分野において，日中間で相関係数の差が大きくなっている．この主たる理由として，中国の「テレビ関連製品」分野においては，「Hisense」，「Skyworth」，「Xiaomi」に代表される会社のように，ネット上の検索において大きな関心を持たれることが相対的に少ないにも関わらず，「テレビ関連製品」分野においては非常に高い知名度があり，通販サイトにおいては高い関心を持たれ，また実際の市場シェアも十分に高い会社が存在しており，主としてこれらの会社が関係する統計部分において相関係数が低下する傾向が観測された．この現象は，現時点では中国特有の現象であると言える．

6. 関連研究

本論文に関する関連研究として，Twitter，検索エンジンの検索数，ブログ，Wikipedia の閲覧数等の情報に基づいて，実社会の動きを予測する手法 [那須野 14, 荒牧 11, Bollen 11, Asur 10, 保住 14, Moat 13] が提案されている． 一方，本論文では，実社会の動きを予測するための情報源として，検索エンジン・サジェストを用いる手法を提案している．

7. おわりに

本論文では，検索における関心の度合いが，通販サイトにおけるページビュー統計との間でどの程度の相関を持つのかについて分析を行った．特に，本論文では，日中二言語を対象として本手法を適用し，日中両言語において，一定以上の相関を示すという結果を報告した．

参考文献

[荒牧 11] 荒牧 英治, 増川 佐知子, 森田 瑞樹：Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測, 情報処理学会研究報告, Vol. 2011–NL–201, (2011)

[Asur 10] Asur, S. and Huberman, B. A.: Predicting the Future with Social Media, in *Proc. WI-IAT*, pp. 492–499 (2010)

[Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003)

[Bollen 11] Bollen, J., Mao, H., and Zeng, X.: Twitter Mood Predicts the Stock Market, *Journal of Computational Science*, Vol. 2, No. 1, pp. 1–8 (2011)

[保住 14] 保住 純, 飯塚 修平, 中山 浩太郎, 高須 正和, 鳴田 絵理子, 須賀 千鶴, 西山 圭太, 松尾 豊：Web マイニングを用いたコンテンツ消費トレンド予測システム, 人工知能学会論文誌, Vol. 29, No. 5, pp. 449–459 (2014)

[今田 16] 今田 貴和, 井上 祐輔, 陳 磊, 徐 凌寒, 宇津呂 武仁, 河田 容英：企業名に関する検索エンジン・サジェストおよびトピックモデリングを用いた市場シェアの分析, 言語処理学会第 22 回年次大会論文集 (2016)

[Moat 13] Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., and Preis, T.: Quantifying Wikipedia Usage Patterns before Stock Market Moves, *Scientific Reports*, Vol. 3, No. 1801 (2013)

[那須野 14] 那須野 薫, 松尾 豊：Twitter における候補者の情報拡散に着目した国政選挙当選者予測, 第 28 回人工知能学会全国大会論文集 (2014)

*5 https://www.taobao.com/

*6 会社単位のページビュー統計である「淘宝」指数 (http://shu.taobao.com/) および各製品分野の検索割合の情報を利用することによって算出した．